Better Computing for Better Bioinformatics

THE WORLD'S FIRST HYBRID-CORE COMPUTER.

CONVEY computer™

George Vacek, PhD, MBA
Director, Convey Life Sciences
gvacek@conveycomputer.com
www.conveycomputer.com/lifesciences/

Acknowledgements

- **Convey**
  - Kirby Collins
  - Wesley Hart
  - Mark Kelly
  - David Soper
  - John Amelio
  - Mike D'Jamoos
  - Mike Ruff
  - Tony Brewer
- **BWA**
  - Heng Li, The Broad
- **Velvet**
  - Daniel Zerbino, UCSC
- **Virginia Bioinformatics Institute**
  - Bob Settlage
  - Lauren McGiver
- **Joint Genome Institute**
  - Alex Copeland
  - Alex Sczyrba
  - Zhong Wang
- **Monsanto Company**
  - Yili Chen
  - Xuefeng Zhou
- **The Jackson Laboratory**
  - Glen Beane

CONVEY computer™

Agenda

- **Better Bioinformatics**
  - High Performance *de novo* Assembly
  - Screening Reads Instead of Contigs
  - High Throughput Resequencing
- **Better Computing**
  - Convey Computers
  - Hybrid-Core Computing

Convey's Hybrid-Core Server Delivers

- **Higher Performance**
  - 5x to 25x application gains
- **Energy Saving**
  - Up to 90% power reduction
- **Easy to use, program, manage**
  - Standard Linux ecosystem
  - Management / Scheduling
  - Programming environment

"Speed and power consumption were our top reasons for selecting the Convey system."

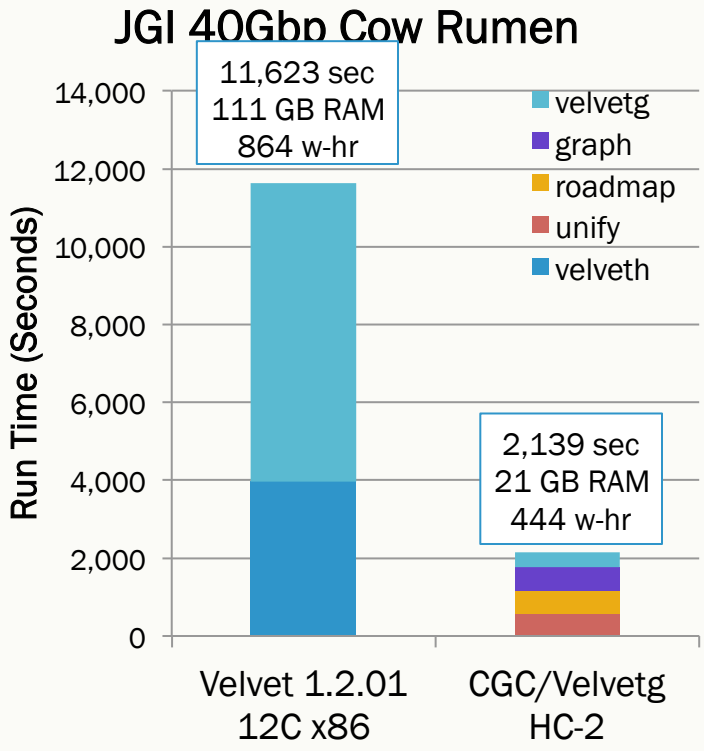
Dr. Guilherme Oliveira, Director Center for Excellence in Bioinformatics

CONVEY computer™


CONVEY
computer™
HIGH PERFORMANCE DE NOVO ASSEMBLY

Reduced Memory Usage, Accelerated Performance - Enables Large Genomes

- **5.4x speed up depends on**
  - Data set size
  - Kmer space complexity
- **RAM reduced 79%**
  - Data types / structures
  - Automated roadmap partitioning
- **1.9x Power Performance**



HC-2: 2 Intel X5670 2.93GHz processors (12 cores total), stripe 4 @ 600GB SATA disks 96GB DDR3 (host), 16GB SG (coprocessor)
X86: host only

"Convey's GraphConstructor offers a new approach to help researchers … to achieve better assemblies or look at bigger jobs such as metagenomic or mammalian genome samples"
Daniel Zerbino, author of Velvet

CONVEY computer™

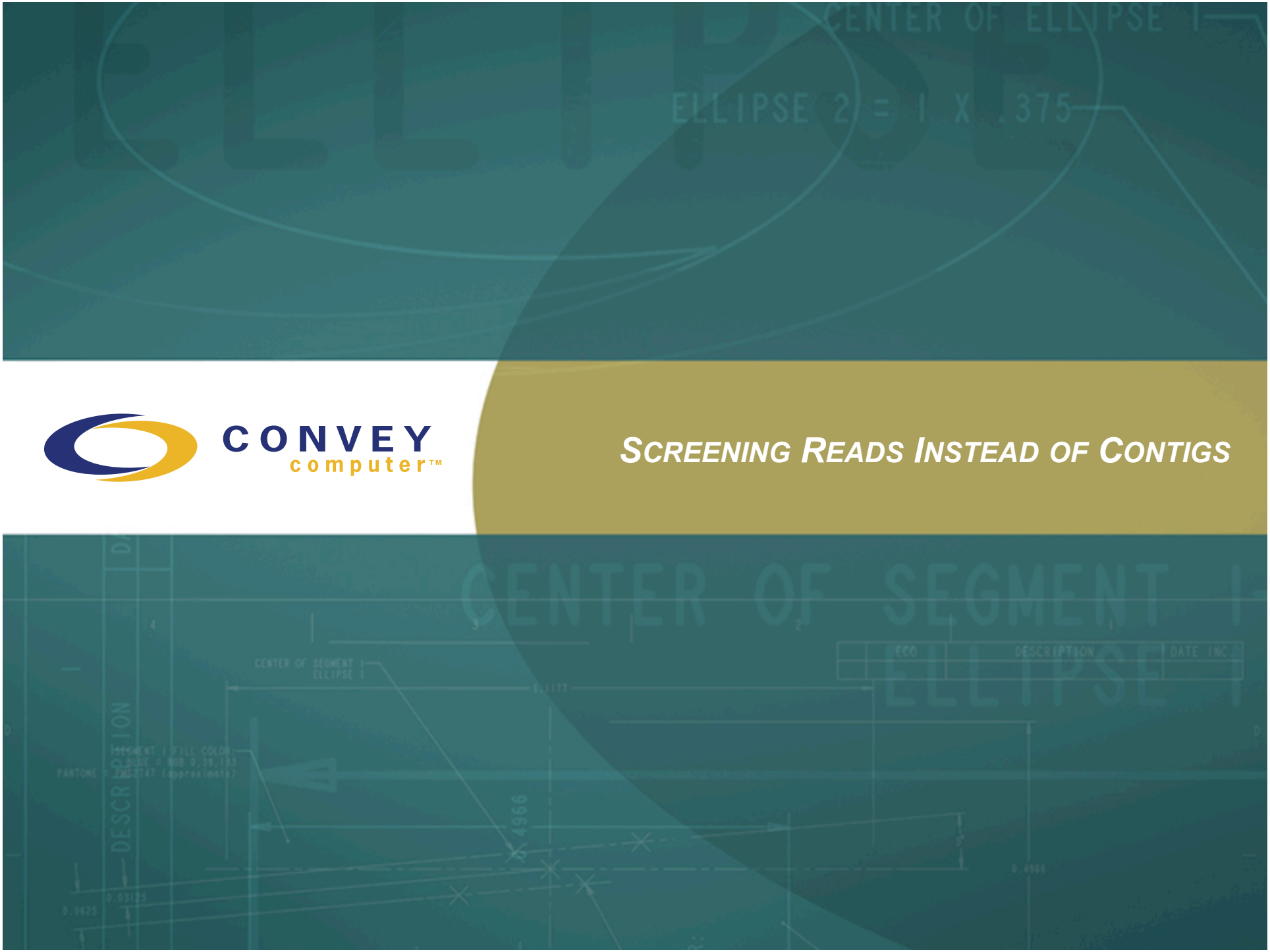
Convey GraphConstructor for *de novo* Assembly

- **Tackle previously impractical genomes**
- **Higher quality assemblies**
- **Lower cost**

- **Interface for Velvet/Oases**
- **Stability, ease of use, optimized workflow**

- **Very fast Kmer Counter**
  - parameter optimization based on roadmap statistics
  - select best kmer length and coverage cutoff

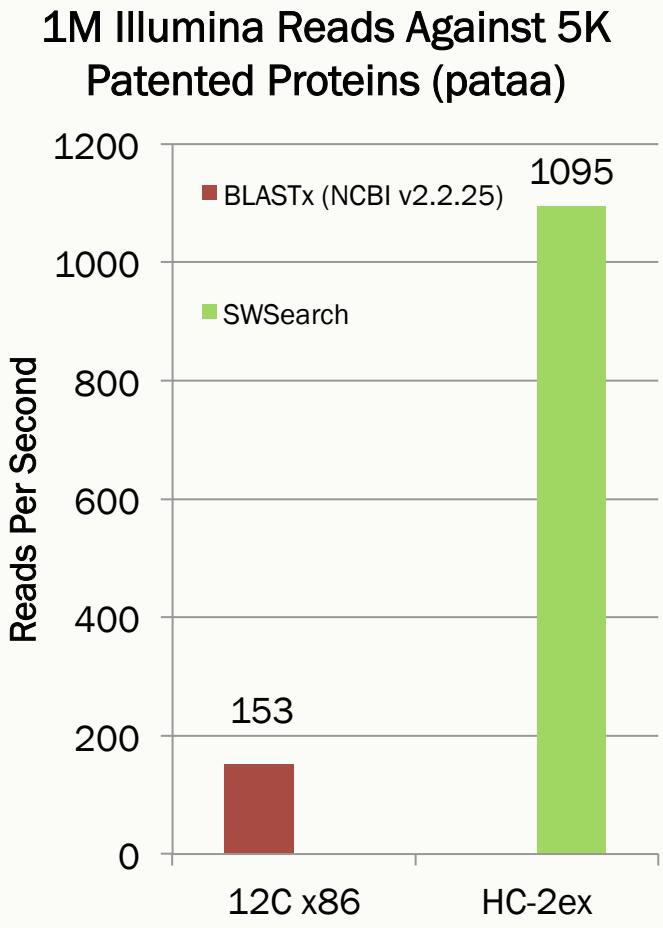
> ***"Convey is solving a big problem here – de novo assembly has been very difficult… Convey has made a significant accomplishment!"***
>
> ***Dr. John Castle, head of Bioinformatics/Genomics, University of Mainz, TrOn***


CONVEY
computer™
SCREENING READS INSTEAD OF CONTIGS

Quickly Identify Reads Associated with Proteins of Interest

- **Translated Search with Smith-Waterman**
- **7.2x faster than BLASTx**
- **2.9x more matches**
  - SWSearch 1081219 hits
  - BLASTx 372344 hits
    - BLAST heuristic filter

```
        19                              23
query:  AARTPKPTAPDSPEMMRG      query:  PGGTLFSSSP
        ::: : :: : ::   ::              :::::::  :
sbjct:  AARLPAPTGPGSPFAGRG      sbjct:  PGGTLFSTXP
        161                             13
```



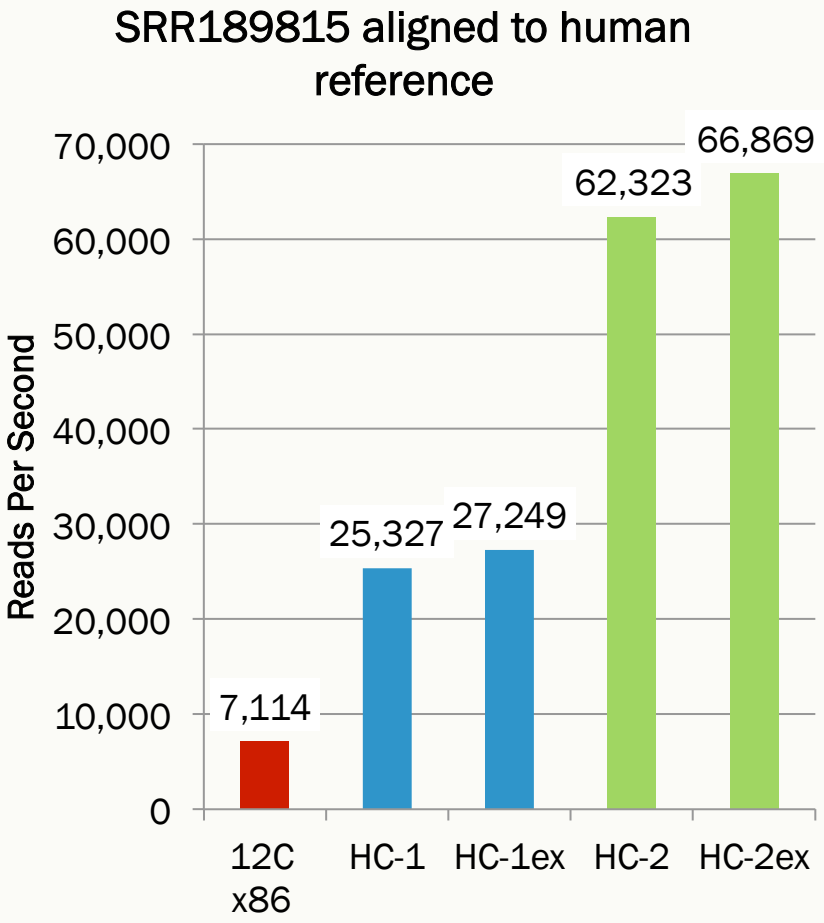
HC-2ex: 192GB (host), 64GB (coproc), stripe 4 @ 600GB SATA
Dell r610: 2 Intel X5680 3.33GHz processors (12 cores total), 96GB of 1333MHz DDR3 memory, stripe 3 @ 146GB SAS

CONVEY computer™


CONVEY
computer™
HIGH THROUGHPUT RESEQUENCING

Workflow Performance for Human

- **BWA 0.5.10 workflow**
  - 2 @ aln + sampe
  - 8.8x - 9.4x over x86
  - 62 - 67 K Reads/Sec
- **Reference G1k v37**
  - 3.1 G bases
- **Reads**
  - HG00124 SRR189815_(1,2)
  - 242 M reads, ~100 bp
  - 24.7 G bases



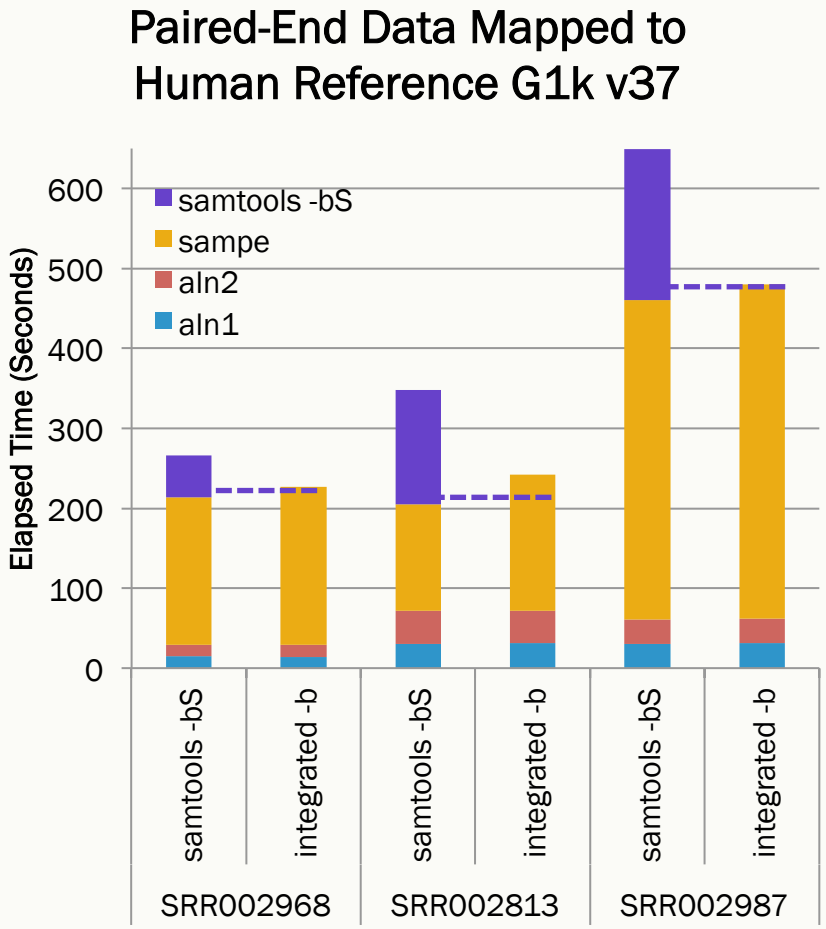
X86: host only from HC-2ex; Intel X5670 2.93GHz processors (12 cores total), stripe 4 @ 600GB SATA disks
HC-1: 128GB (host), 64GB (coproc), stripe 2 @ 1TB SATA disks
HC-1ex: 128GB (host), 64GB (coproc), stripe 2 @ 1TB SATA
HC-2: 96GB DDR3 (host), 16GB SG (coprocessor)
HC-2ex: 192GB DDR3 (host), 64GB SG (coprocessor)

CONVEY computer™

Further Workflow Optimization
Integrated BAM format generation

- **SAM to compressed BAM**
  - samtools *vs.* integrated sampe
- **3.9 - 9.8x speed up**
- **Even greater savings for slow file systems**

- **Reference G1k v37**
  - 3.1 G bases
- **Paired-end Reads**
  - SRR002813, SRR002987, SRR002968



HC-2ex: 192GB DDR3 (host), 64GB SG (coprocessor), stripe 4 @ 600GB SATA disks

CONVEY computer™

The Jackson Laboratory

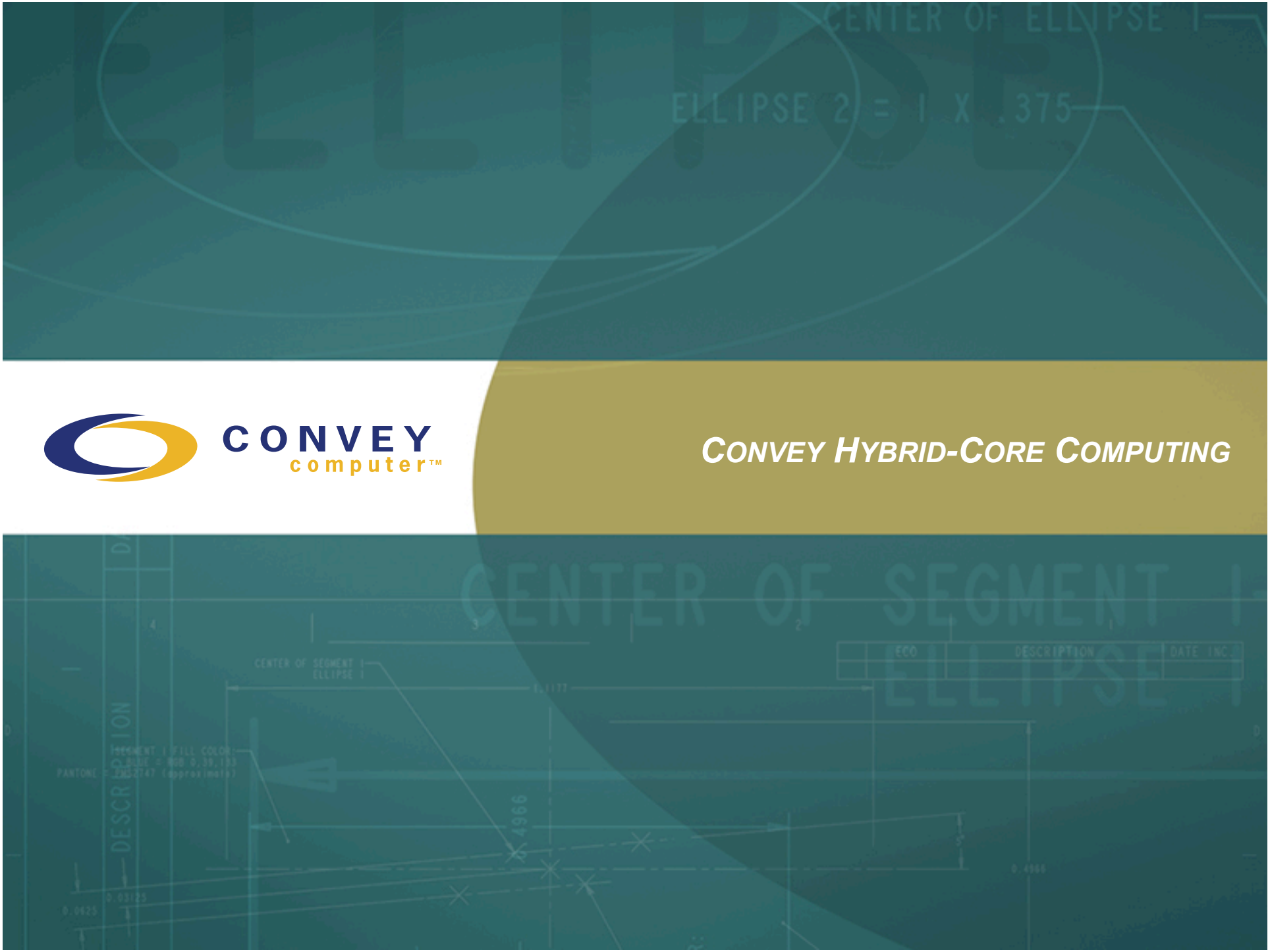
- **Mutations vary in size**
  - E.g. translocation breakpoint
  - Want reads that span breakpoint
- **Run BWA with varying parameters**
  - Get more of these mutations
- **Too slow on 32-core servers**
- **HC-2ex is 11.3x faster**
  - Afford to adjust parameters
  - Quickly perform multiple runs
  - Achieve better results

"We found GPUs weren't a good fit for alignment… the performance isn't that compelling. Other FPGA system vendors didn't have the number of tools Convey does or the system wasn't as easy to use. Also a developer community is evolving around the Convey systems where we could share third-party tools."

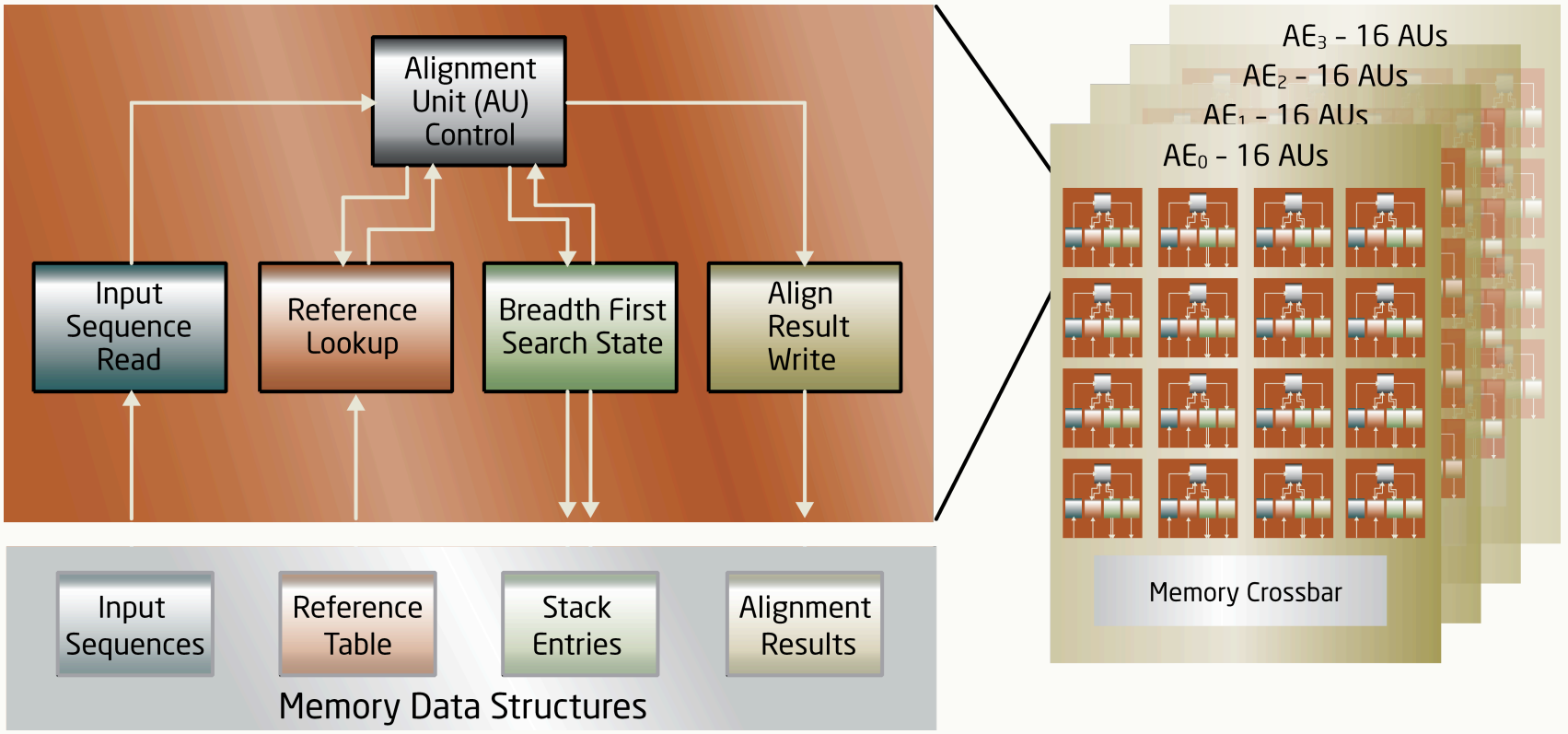
Glen Beane
The Jackson Laboratory

BWA 0.5.10
X86: 4 x 8-core AMD Magny Cours 2.4GHz Opteron
HC-2ex: 2 x 6-core Intel X5670 2.93GHz, coprocessor

CONVEY computer™


CONVEY
computer™
CONVEY HYBRID-CORE COMPUTING

BWA Personality



- Implemented in hardware on coprocessor FPGAs
- Highly parallel—up to 2,048 simultaneous alignment operations
  - 64 alignment units each operate on 32 sequences simultaneously
- Leverages Convey HC highly parallel memory

CONVEY computer™

Bioinformatics Applications on HC

	Organization	Application	Usage
Convey Bioinf Suite	Convey Bioinformatics Suite	BWA	Reference Mapping
	Convey Bioinformatics Suite	Velvet/CGC	*De Novo* Assembly
	Convey Bioinformatics Suite	Kmer Counter	Read Analysis for Assembly
	Convey Bioinformatics Suite	SWSearch	Smith-Waterman Search
	Convey Bioinformatics Suite	BLAST(p,x)	Protein Database Search
Available	CLC bio	CLC Genomics platform	Analysis, workflows, visualization
	Michigan Technological University	PCAP	Overlap-based Assembly
	University of California San Diego	InsPect	Protein Assembly with PTMs
Performance Proven	BlueSpec	Memocode	Burrows-Wheeler Aligner
	Iowa State University	RMAP, Shepard	Short-read Mapping
	Technical University Crete	BLASTn	Nucleotide Sequence Search
	University of California Los Angeles	Fluid Registration	Medical Imaging
	University of California Riverside	BowTie/FHAST	Burrows-Wheeler Aligner
	University of South Carolina	Mr Bayes	Phylogenetics
Project Initiated	Boston University	BLAST(p,x)	Protein Database Search
	Free University of Berlin	SeqAn	Sequence Analysis Library
	Technical University Darmstadt	GROMACS	Molecular Dynamics
	Bielefeld University	SARUMAN	Short-read Mapping
	University of Paderborn	Suffix Tree	Short-read Mapping
	University of Washington	BFAST	Short-read Mapping
	Virginia Bioinformatics Institute	Various	Mol Dynamics, Bioinformatics

High Throughput Bioinformatics

- **In-house development and collaborations**
  - Customers and partners
  - Software vendors
  - Instrument manufacturers
  - Cloud services
- **Addressing many facets of bioinformatics**
  - primary analysis
  - de novo assembly and reference mapping
  - sequence alignment and search
  - annotation, other downstream analysis
- **www.conveycomputer.com/lifesciences/**


CONVEY
computer™
THANK YOU


CONVEY
computer™
END